



贝叶斯统计学

Bayesian Statistics

开课的话

- 历史悠久：R. T. Bayes(1701-1761)
P. C. Laplace(1749-1827)
- 争论不休：经典学派**VS**贝叶斯学派
- 困难所在：模型复杂，计算量巨大
- 应用广泛：不但在统计本身而且在许多其它学科上
都有重要应用
- 欣欣向荣：电子计算机；算法；近二十多年来大发展




Reverend Thomas Bayes

- 课堂纪律：有病有事一律向系里请假，而不是向我请假。有系里批准的假条给我，我都没异议。每次上课都点名，出勤率关系到你的成绩。
- 学习态度：强烈的求知（非求职）欲望。
- 作业：每次作业都有登记评分，另有贝叶斯统计英译中作业（**12月31日**完成上交。期末考试将有英语题）。
- 问与答：没有愚蠢的问题，只有愚蠢的回答。任何问题都可向我提出，我会尽自己的能力，回答你们的问题。如果没有提问，则认定你已经懂了所教内容。

第一章 先验分布与后验分布

1.1 三种信息

一、 总体信息，即总体分布或总体所属分布族给我们的信息，譬如，“总体是正态分布”这一句话就给我们带来很多信息：它的密度函数是一条钟形曲线；它的一切阶矩都存在；有关正态变量（服从正态分布的随机变量）的一些事件的概率可以计算；有关正态分布可以导出 χ^2 分布、t 分布和 F 分布等重要分布；还有许多成熟的点估计、区间估计和假设检验方法可供我们选用。总体信息是很重要的信息，为了获取此种信息往往耗资巨大。




二、样本信息，即从总体抽取的样本给我们提供的信息。这是最“新鲜”的信息，并且愈多愈好。人们希望通过对样本的加工和处理对总体的某些特征做出较为精确的统计推断。没有样本就没有统计学可言。这是大家都理解的事实。

- 基于上述两种信息进行的统计推断被称为经典统计学，它的基本观点是把数据（样本）看成是来自具有一定概率分布的总体，所研究的对象是这个总体而不局限于数据本身。

三、**先验信息**，即在抽样之前有关统计问题的一些信息，一般说来，先验信息主要来源于经验和历史资料。先验信息在日常生活和工作中也经常可见，不少人在自觉地或不自觉地使用它。看下面二个例子。

例1.1 英国统计学家**Savage(1961)**曾考察如下二个统计实验：

- **A.** 一位常饮牛奶的妇女声称，她能辨别先倒进杯子里的是茶还是牛奶。对此做了十次试验，她都正确地说出了。
- **B.** 一位音乐家声称，他能从一页乐谱辨别出是海邓（**Haydn**）还是莫扎特（**Mozart**）的作品。在十次这样的试验中，他都能正确辨别。




在这两个统计试验中，假如认为被实验者是在猜测，每次成功的概率为**0.5**，那么十次都猜中的概率为 $2^{-10} = 0.0009766$ ，这是一个很小的概率，是几乎不可能发生的，所以“每次成功概率为**0.5**”的假设应被拒绝。被实验者每次成功概率要比**0.5**大得多。这就不是猜测，而是他们的经验在帮他们的忙。可见经验（先验信息的一种）在推断中不可忽视，应加以利用。


例 1.2 “免检产品”是怎样决定的？某厂的产品每天都要抽检几件，获得不合格率 θ 的估计。经过一段时间后就积累大量的资料，根据这些历史资料（先验信息的一种）对过去产品的不合格率可构造一个分布：

$$P(\theta = \frac{i}{n}) = \pi_i, \quad i = 0, 1, \cdots n$$

这个对先验信息进行加工获得的分布今后称为先验分布。这个先验分布是综合了该厂过去产品的质量情况。如果这个分布的该率绝大部分集中在 $\theta = 0$ 附近，那该产品可认为是“信得过产品”。假如以后的多次抽检结果与历史资料提供的先验分布是一致的。使用单位就可以对它做出“免检产品”的决定，或者每月抽检一、二次就足够了，这就省去了大量的人力与物力。可见历史资料在统计推断中应加以利用。



基于上述三种信息（**总体信息**、**样本信息**和**先验信息**）进行的统计推断被称为贝叶斯统计学。它与经典统计学的主要差别在于是否利用先验信息。在使用样本信息上也是有差异的。贝叶斯学派重视已出现的样本观察值，而对尚未发生的样本观察值不予考虑，贝叶斯学派很重视先验收集、挖掘和加工，使它数量化，形成先验分布，参加到统计推断中来，以提高统计推断的质量。忽视先验信息的利用，有时是一种浪费，有时还会导致不合理的结论。



贝叶斯学派的最基本的观点是：任一个未知量 θ 都可看作一个随机变量，应用一个概率分布去描述对 θ 的未知状况。这个概率分布是在抽样前就有的关于 θ 的先验信息的概率陈述。这个概率分布被称为先验分布、有时还简称为先验（Prior）。因为任一未知量都有不确定性，而在表述不确定性程度时，概率与概率分布是量好的语言。例 1.2 中产品不合格率 θ 是未知量，但每天都有一些变化，把它看作一个随机变量是合适的，用一个概率分布去描述它也是很恰当的。即使是一个几乎不变的未知量，用一个概率分布去描述它的不确定性也十分合理的。

例1.3 学生估计一新教师的年龄。依据学生们的生活经历。在看了新教师的照片后立即会有反应：“新教师的年龄在**30岁到50岁**之间，极有可能在**40岁**左右。”一位统计学家与学生们交谈，明确这句话中“左右”可理解为岁，“极有可能”可理解为**90%**的把握。于是学生们对新教师年龄（未知量）的认识（先验信息）可综合为图1.1所示的概率分布，这也是学生们对未知量（新教师年龄）的概率表述。

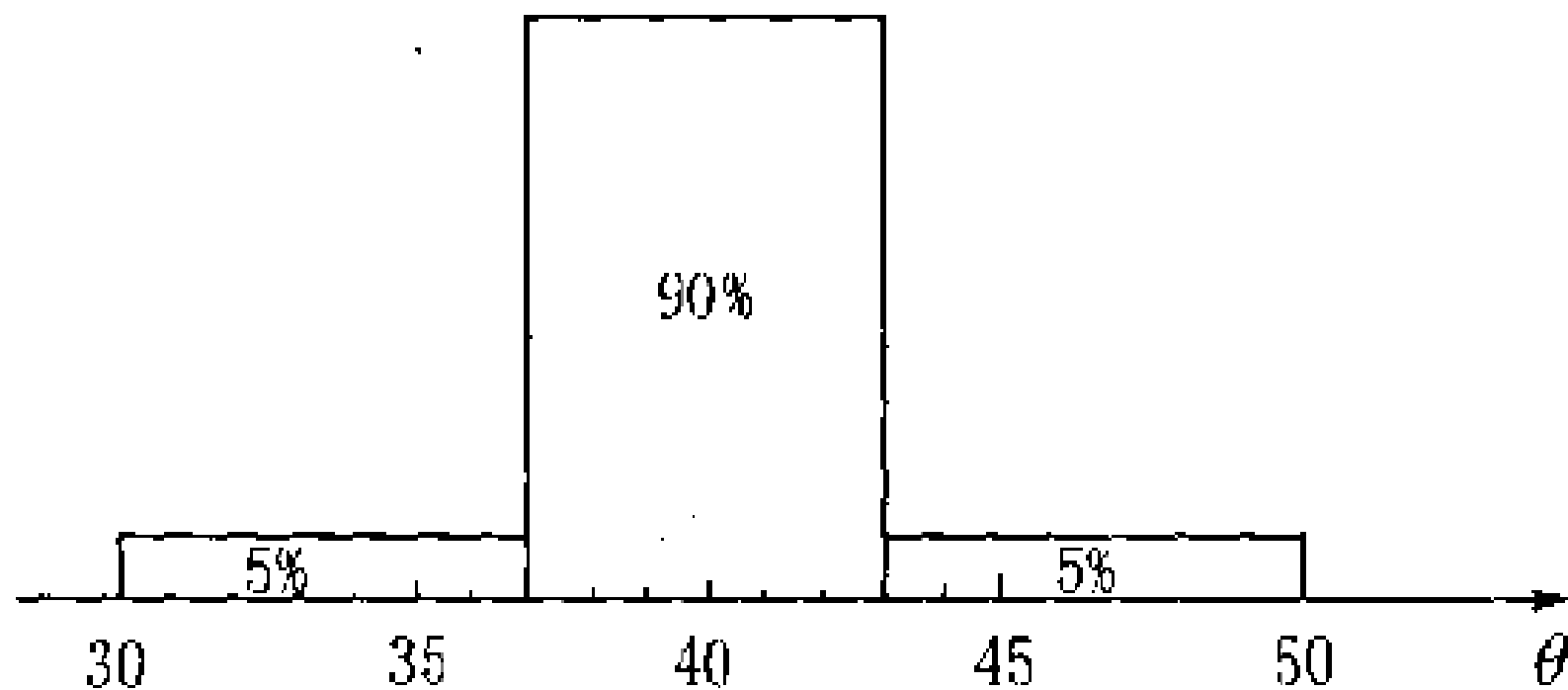




图 1.1 新教师年龄的先验分布



第一，按图 1.1 所示的概率分布我们可谈论未知量 θ 位于某个区间的概率。譬如， θ 位于 37 到 43 岁间的概率为 0.90，即

$$P(37 \leq \theta \leq 43) = 0.90$$

可这个概率陈述在经典统计中是不允许的，因为经典统计认为 θ 是常量，它要么是 37 到 43 岁之间（概率为 1），要么在这个区间之外（上述事件概率为零 0），不应有 0.9 的概率。可在实际中类似的说法经常听到。譬如“某逃犯的年龄大约 35 岁左右”、“明日降水概率为 0.85”、“某学生能考上大学的概率为 0.95”、“这场足球赛甲队能胜的概率只有 0.6 左右”。这样的概率陈述能为大多数人理解、接受和采用。



第二，图1.1中的概率0.90不是在大量重复试验中获得的。而是学生们根据自己的生活经历的积累对该事件发生可能性所给出的信念，这样给出的概率在贝叶斯统计中是允许的，并称为主观概率。它与古典概率和用频率确定的概率有相同的含义，只要它符合概率的三条公理即可。贝叶斯学派认为引入主观概及由此确定的先验分布至少把概率与统计的研究与应用范围扩大到不能大量重复的随机现象中来。其次，主观概率的确定不是随意的，而是要求当事人对所观察的事件有较透彻的了解和丰富的经验，甚至是这一行的专家，在这个基础上确定的主观概率就能符合实际。

1.2 贝叶斯公式(定理)

贝叶斯方法是基于贝叶斯定理而发展起来用于系统地阐述和解决统计问题的方法。

一、贝叶斯公式的密度函数形式

1. 依赖于参数 θ 的密度函数在经典统计中记为 $p(x;\theta)$ 或 $p_{\theta}(x)$ ，它表示在参数空间 $\Theta = \{\theta\}$ 中不同的 θ 对应不同的分布。可在贝叶斯统计中记为 $p(x|\theta)$ ，它表示在随机变量 θ 给定某个值时，总体指标 X 的条件分布。

2. 根据参数 θ 的先验信息确定先验分布 $\pi(\theta)$ 。

3. 从贝叶斯观点看，样本 $x = (x_1, \dots, x_n)$ 的产生要分二步进行。首先设想从先验分布 $\pi(\theta)$ 产生一个样本 θ ，第二步是从总体分布 $p(x|\theta)$ 产生一个样本 $x = (x_1, \dots, x_n)$ ，这个样本是具体的，人们能看得到的，此样本 x 发生概率是与如下联合密度函数成正比：

$$p(x|\theta') = \prod_{i=1}^n p(x_i|\theta')$$

这个联合密度函数是综合了总体信息和样本信息，常称为似然函数，记为 $L(\theta')$ 。频率学派和贝叶斯学派都承认似然函数，二派认为：在有了样本观察值 $x = (x_1, \dots, x_n)$ 后，总体和样本中所含 θ 的信息都被包含在似然函数 $L(\theta')$ 之中。

4. 由于 θ' 是设想出来的，它仍然是未知的，它是按先验分布 $\pi(\theta)$ 而产生的，要把先验信息进行综合，不能只考虑 θ' ，而应对 θ 的一切可能加以考虑。故要用 $\pi(\theta)$ 参与进一步综合。这样一来，样本 x 和参数 θ 的联合分布

$$h(x, \theta) = p(x|\theta)\pi(\theta)$$

把三种可用的信息都综合进去。

5. 我们的任务是要对未知数 θ 做出统计推断。在没样本信息时，人们只能据先验分布对 θ 做出推断。在有样本观察值 $x = (x_1, \dots, x_n)$ 之后，我们应该依据 $h(x, \theta)$ 对 θ 做出推断。为此我们需把 $h(x, \theta)$ 作如下分解

$$h(x, \theta) = \pi(x|\theta)m(x)$$

其中 $m(x)$ 是 x 的边缘密度函数。

$$m(x) = \int_{\Theta} h(x, \theta) d\theta = \int_{\Theta} p(x|\theta) \pi(\theta)$$

它与 θ 无关，或者说， $m(x)$ 中不含 θ 的任何信息。因此能用来对 θ 做出推断的仅是条件分布 $\pi(\theta|x)$ 。它的计算公式是

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta} \quad (1.1)$$

这就是贝叶斯公式密度函数形式。这个在样本 x 给定下， θ 的条件分布被称为 θ 的后验分布。它是集中了总体、样本和先验等三种信息中有关 θ 的一切信息，而又是排除一切与 θ 无关信息之后所得到的结果。故基于后验分布 $\pi(\theta|x)$ 对 θ 进行统计推断是更为有效，也是最合理的。

6. 在 θ 是离散随机变量时, 先验分布可用先验分布列 $\pi(\theta)$, $i = 1, 2, \dots$, 表示。这时后验分布也是离散形式。

$$\pi(\theta_i|x) = \frac{p(x|\theta_i)}{\sum_j p(x|\theta_j)\pi(\theta_j)}, i = 1, 2, \dots \quad (1.2)$$

假如总体 X 也是离散的, 那只要把 (1.1) 或 (1.2) 中的密度函数 $p(x|\theta)$ 看作为概率函数 $P(X = x|\theta)$ 即可。

二. 后验分布是三种信息的综合

先验分布 $\pi(\theta)$ 是反映人们在抽样前对 θ 的认识, 后验分布 $\pi(\theta|x)$ 是反映人们在抽样后对 θ 的认识。之间的差异是由于样本 x 出现后人们对 θ 认识的一种调整。所以后验分布 $\pi(\theta|x)$ 可以看作是人们用总体信息和样本信息（综合称为抽样信息）对先验分布 $\pi(\theta)$ 作调整的结果。

例 1.4 设事件 A 的概率为 θ ，即 $\pi(A) = \theta$ 。为了估计 θ 而作 n 次独立观察，其中事件 A 出现次数为 X ，显然， X 服从二项分布 $b(n, \theta)$ ，即

$$P(X = x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, x = 0, 1, \dots, n.$$

这就是似然函数。假如在试验前我们对事件 A 没有什么了解，从而对其发生概率 θ 也说不出是大是小。在这种场合，用均匀分布 $U(0, 1)$ 作为 θ 的先验分布。这时 θ 的先验分布为

$$\pi(\theta) = \begin{cases} 1, & 0 < \theta < 1 \\ 0, & \text{其它场合} \end{cases} \quad (1.3)$$

为了综合抽样信息和先验信息，可利用贝叶斯公式，为此计算样本 \mathbf{X} 与参数 θ 的联合分布

$$h(x, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, x = 0, 1, \dots, n, 0 < \theta < 1.$$

此式在定义域上与二项分布有差别。再计算 \mathbf{X} 的边缘分布

$$m(x) = \int_0^1 h(x, \theta) d\theta = \binom{n}{x} \int_0^1 \theta^x (1 - \theta)^{n-x} d\theta = \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} = \frac{1}{n+1},$$

$x = 0, 1, \dots, n.$

最后可得 θ 的后验分布

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \theta^{(x+1)-1} (1 - \theta)^{(n-x+1)-1}, 0 < \theta < 1$$

这个分布不是别的，就是贝塔分布 $\text{Be}(x+1, n-x+1)$ 。

注：两个著名函数

- 贝塔函数 $\beta(z, w) = \int_0^1 t^{z-1} (1-t)^{w-1} dt$

- 伽玛函数 $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$

- 性质

$$\Gamma(z+1) = z\Gamma(z)$$

$$\beta(z, w) = \frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)}$$

拉普拉斯在 1786 研究了巴黎的男婴出生的比率，他希望检验男婴出生的概率 θ 是否大于 0.5。为此他收集到 1745 年到 1770 年在巴黎出生的婴儿数据。于是可得男婴 251527 个，女婴 241945 个。他先用 $U(0, 1)$ 作为 θ 的先验分布。于可得 θ 的后验分布 $Be(x+1, n-x+1)$ ，其中 $n=251527+241945=493472$ ， $x=251527$ 。利用这个后验分布，拉普拉斯计算了 “ $\theta \leq 0.5$ ” 的后验概率

$$P(\theta \leq 0.5|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \int_0^{0.5} \theta^x (1-\theta)^{n-x} d\theta$$

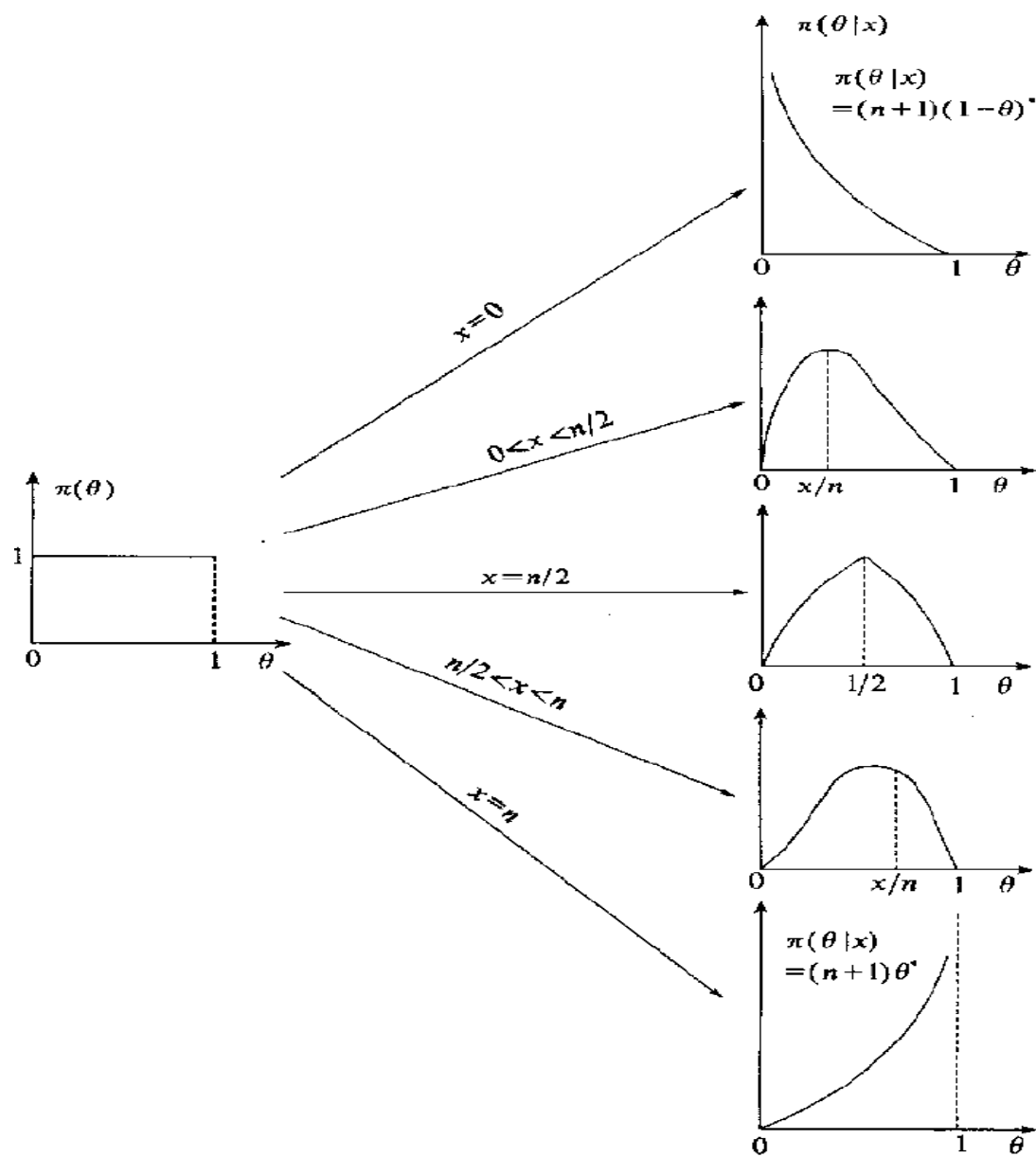


图 1.2 抽样信息对先验分布的调整

1.3 共轭先验分布

一、 共轭先验分布

大家都知道，在区间 $(0, 1)$ 上的均匀分布是贝塔分布 $\text{Be}(1, 1)$ 。这时从例 1.4 中可以看到一个有趣的现象。二项分布 $b(n, \theta)$ 中的成功概率 θ 的先验分布若取 $\text{Be}(1, 1)$ ，则其后验分布也是贝塔分布 $\text{Be}(x+1, n-x+1)$ 。其中 x 为 n 次独立试验中成功出现次数。先验分布与后验分布同属于一个贝塔分布族，只是其参数不同而已。这一现象不是偶然的，假如把 θ 的先验分布换成一般的贝塔分布 $\text{BE}(\alpha, \beta)$ ，其中 $\alpha > 0, \beta > 0$ 。经过类似计算可以看出， θ 的后验分布面仍是贝塔分布 $\text{BE}(\alpha+x, \beta+n-x)$ ，此种先验分布被称为 θ 的共轭先验分布。

定义 1.1 设 θ 是总体分布中的参数（或参数向量）， $\pi(\theta)$ 是 θ 的先验密度函数，假如由抽样信息算得的后验密度函数与 $\pi(\theta)$ 有相同的函数形式，则称 $\pi(\theta)$ 是 θ 的（自然）共轭先验分布。

例1.6 正态均值（方差已知）的共轭先验分布是正态分布。
设 x_1, \dots, x_n 是来自正态分布 $N(\theta, \sigma^2)$ 的一个样本观值。其中 σ^2 已知。
此样本的似然函数为

$$P(x|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right\},$$
$$-\infty < x_1, \dots, x_n < +\infty \quad (1.5)$$

现取另一个正态分布 $N(\mu, \tau^2)$ 作为正态均值 θ 的先验分布, 即

$$\pi(\theta) = \left(\frac{1}{\sqrt{2\pi}\tau}\right) \exp\left\{-\frac{(\theta - \mu)^2}{2\tau^2}\right\}, -\infty < \theta < +\infty \quad (1.6)$$

其中 μ 与 τ^2 为已知, 由此可以写出样本 x 与参数 θ 的联合密度函数

$$h(x|\theta) = k_1 \exp\left\{-\frac{1}{2}\left[\frac{n\theta^2 - 2n\theta\bar{x} + \sum_{i=1}^n x_i^2}{\sigma^2} + \frac{\theta^2 - 2\mu\theta + \mu^2}{\tau^2}\right]\right\}$$

其中 $k_1 = (2\pi)^{-(n+1)/2} \tau^{-1} \sigma^{-n}$, $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$ 。若再记

$$\sigma_0^2 = \frac{\sigma^2}{n}, A = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2}, B = \frac{\bar{x}}{\sigma_0^2} + \frac{\mu}{\tau^2}, C = \frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu^2}{\tau^2}$$

则有

$$h(x, \theta) = k_1 \exp\left\{-\frac{1}{2}[A\theta^2 - 2\theta B + c]\right\} = k_2 \exp\left\{-\frac{(\theta - B/A)^2}{2/A}\right\}$$

其中 $k_2 = k_1 \exp\left\{-\frac{1}{2}(c - B^2/A)\right\}$ 。由此容量算得样本 x 的边缘分布

$$m(x) = \int_{-\infty}^{\infty} h(x, \theta) d\theta = k_2 \left(\frac{2\pi}{A}\right)^{\frac{1}{2}}$$

上面两式相除，即得 θ 的后验分布

$$\pi(\theta|x) = \left(\frac{2\pi}{A}\right)^{\frac{1}{2}} \exp\left\{-\frac{(\theta - B/A)^2}{2/A}\right\} \quad (1.7)$$

这是正态分布，其均值 μ_1 与方差 τ_1^2 分别为

$$\mu_1 = \frac{B}{A} = \frac{\bar{x}\sigma_0^{-2} + \mu\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}}, \quad \frac{1}{\tau^2} = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2} \quad (1.8)$$

这就说明了正态均值（方差已知）的共轭先验分布是

正态分布。如设 $X \sim N(\theta, 2^2)$, $\theta \sim N(10, 3^2)$ 。若从正态总

体 \mathbf{X} 抽得容量为 5 的样本，算得 $\bar{x} = 12.1$ ，于是可从 (1.8)

算得 $\mu_1 = 11.93$ 和 $\tau_1^2 = (\frac{6}{7})^2$ 。这时正态均值 θ 的后验分布 $N(11.93, (\frac{6}{7})^2)$

二、 后验分布的计算

在给定样本分布 $p(x|\theta)$ 和后验分布 $\pi(\theta)$ 后可用贝叶斯公式计算 θ 的后验分布

$$\pi(\theta) = p(x|\theta)\pi(\theta) / m(x)$$

由于 $m(x)$ 不依赖于 θ ，在计算 θ 的后验分布中仅起到一正则化因子的作用。假如把 $m(x)$ 省略，把贝叶斯分式改写为如等价形式

$$\pi(\theta) \propto p(x|\theta)\pi(\theta) \quad (1.9)$$

其中符号 “ \propto ” 表示两边仅差一个常数因子，一个不依赖于 θ 的常数因子。(1.9) 式右端虽不是正常的密度函数，但它是后验分布 $\pi(\theta|x)$ 的核，在需要时可以利用适当方式计算出后验密度，特别当看出 $\pi(\theta|x)$ 的核就是某常用分布的核时，不用计算 $m(x)$ 就可很快恢复所缺常数因子。

例1.7 二项分布中的成功概率 θ 的共轭先验分布是贝塔分布。

设总体 $X \sim b(n, \theta)$ ，其密度函数中与 θ 有关部分为 $\theta^x (1-\theta)^{n-x}$ 。又

设 θ 的先验分布为贝塔分布 $Be(\alpha, \beta)$ ，其核为 $\theta^{\alpha-1} (1-\theta)^{\beta-1}$ ，其中

α, β 已知，从而可写出 θ 的后验分布

$$\pi(\theta|x) \propto \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1}, 0 < \theta < 1$$

立即可以看出，这是贝塔分布 $Be(\alpha+x, \beta+n-x)$ 的核，故此 posterior 密度为

$$\pi(\theta|x) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + x)\Gamma(\beta + n - x)} \theta^{\alpha+x-1} (1-\theta)^{\beta+n-x-1}, 0 < \theta < 1$$

一、 共轭先验分布的优秀缺点

共轭先验分布在很多场合被采用，因为它有两个优点：

1. 计算方便；
2. 后验分布的一些参数可得到很好的解释。

例 1.8 在“正态均值 θ 的共轭先验分布为正态分布”的例 1.6 中，其后验均值 μ_1 （见（1.8）式）可改写为

$$\mu_1 = \frac{\sigma_0^{-2}}{\sigma_0^{-2} + \tau^{-2}} \bar{x} + \frac{\tau^{-2}}{\sigma_0^{-2} + \tau^{-2}} \mu = \gamma \bar{x} + (1 - \gamma) \mu$$

其中 $\gamma = \sigma_0^{-2} / (\sigma_0^{-2} + \tau^{-2})$ 是用方差倒数组成的权，于是
后验均值 μ_1 是样本均值 \bar{x} 与先验均值 μ 的加权平均。

在处理正态分布时，方差的倒数发挥着重要的作用，并称其为精度，于是在正态均值的共轭先验分布的讨论中，其后验方差 τ^2 所满足的等式（见（1.8 式）

$$\frac{1}{\tau_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\tau^2} = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

可解释为：后验分布的精度是样本均值分布的精度与先验分布精度之和，增加样本量 n 或减少先验分布方差都有利于提高后验分布的精度。

例 1.9 在“二项的成功概率 θ 的共轭先验分布是贝塔分布”的例 1.7 中，后验分布 $\text{Be}(\alpha + x, \beta + n - x)$ 的均值与方差亦可改写为

$$E(\theta|x) = \frac{\alpha + x}{\alpha + \beta + n} = \frac{n}{\alpha + \beta + n} \frac{x}{n} + \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} = \gamma \cdot \frac{x}{n} + (1 - \gamma) \cdot \frac{\alpha}{\alpha + \beta}$$

$$\text{Var}(\theta|x) = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta|x)[1 - E(\theta|x)]}{\alpha + \beta + n + 1}$$

其中 $\gamma = n/(\alpha + \beta + n)$ ， x/n 是样本均值， $\alpha/(\alpha + \beta)$ 是先验均值，从上述加权平均可见，后验均值是介于样本均值与先验均值之间，它偏向那一侧由 γ 的大小决定。

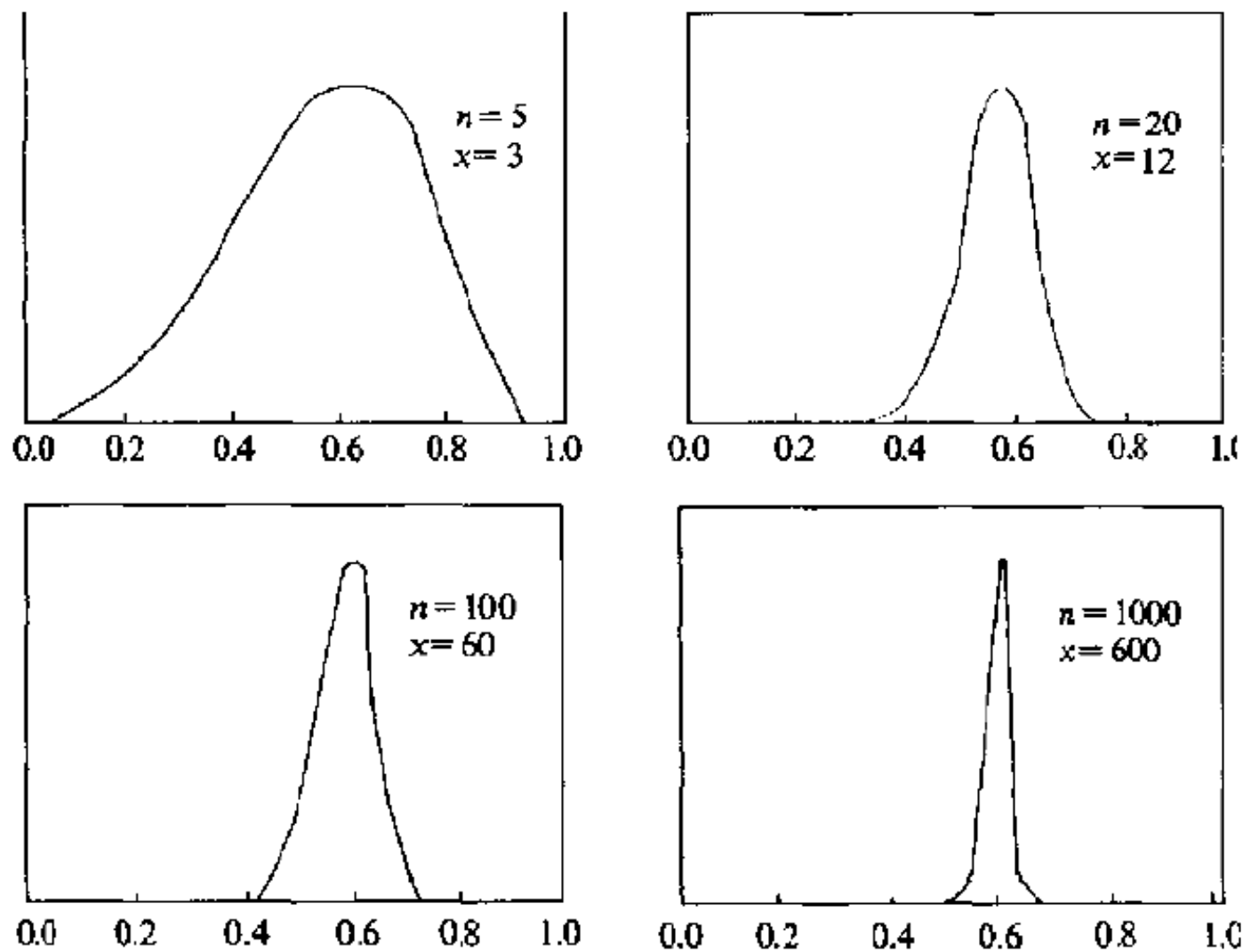


图 1.3 当 n 与 x 成比例增加时,后验密度——贝塔分布
 $Be(\alpha+x, \beta+n-x)$ —变化情况(纵坐标刻度不同)

四、常用的共轭先验分布

共轭先验分布的选取是由似然函数 $L(\theta) = p(x|\theta)$ 中所含 θ 的因式所决定的，即选取与似然函数（ θ 的函数）具有相同核的分布作为先验分布。若此想法得以实现，那共轭先验分布就产生了。

例 1.10 设 x_1, \dots, x_n 是来自正态分布 $N(\theta, \sigma^2)$ 的样本观测值，其中 θ 已知。现要寻求 σ^2 的共轭先验分布。样本的似然函数为

$$p(x|\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right\} \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right\}$$

上述似然函数中 σ^2 的因式将决定 σ^2 的共轭先验分布的形式，什么分布具有上述的核呢？

设 X 服从伽玛分布 $\text{Ga}(\alpha, \lambda)$, 其中 $\alpha > 0$ 为形状参数, $\lambda > 0$ 为尺度参数, 其密度函数为

$$p(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0$$

通过概率运算可以求得 $Y = X^{-1}$ 的密度函数

$$p(y|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{y}\right)^{\alpha+1} e^{-\frac{\lambda}{y}}, y > 0$$

这个分布称为倒伽玛分布, 记为 $\text{IGa}(\alpha, \lambda)$, 假如取此倒伽玛分布为 σ^2 的先验分布, 其中参数 α 与 λ 为已知, 则其密度函数为

$$\pi(\sigma^2) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} e^{-\lambda/\sigma^2}, \sigma^2 > 0$$

于是 σ^2 的后验分布为

$$\begin{aligned}\pi(\sigma^2|x) &\propto p(x|\sigma^2)\pi(\sigma^2) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\alpha+\frac{n}{2}+1} \exp\left\{-\frac{1}{\sigma^2}\left[\lambda + \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right]\right\}\end{aligned}$$

容易看出，这仍是倒伽玛分布 $IGa(\alpha + \frac{n}{2}, \lambda + \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2)$ ，它

是正态方差 σ^2 的共轭先验分布，其合理性由先验信息决定。

1.5 多参数模型

统计中很多实际问题含有多个未知参数，如正态总体 $N(\mu, \sigma^2)$ 常含有二个未知参数 μ 与 σ^2 ，又如多项分布 $M(n; p_1, \dots, p_k)$ 常含有 $k-1$ 个未知参数，至于多元正态分布 $N(\mu, \Sigma)$ 则含有更多具参数。

在贝叶斯方法的框架中处理多参数的方法与处理单参数方法相似，先根据先验信息给出参数的先验分布，然后按贝叶斯公式算得后验分布，为确定起见，设总体只含二个参数 $\theta = (\theta_1, \theta_2)$ ，总体的密度函数为 $p = (x | \theta_1, \theta_2)$ ，若从该总体抽取一个样本 $x = (x_1, \dots, x_n)$ ，并给出先验密度 $\pi = (\theta_1, \theta_2)$ ，则 (θ_1, θ_2) 的后验密度为

$$\pi(\theta_1, \theta_2 | x) \propto p(x | \theta_1, \theta_2) \pi(\theta_1, \theta_2) \quad (1.9)$$

例1.12 正态均值与正态方差的（联合）共轭先验分布。设 x_1, \dots, x_n 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本，该样本的联合密度函数为

$$\begin{aligned} p(x|\mu, \sigma^2) &\propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{x} - \mu)^2]\right\} \end{aligned}$$

其中 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$ ，以下寻找 (μ, σ^2) 的共轭先验分布。

考虑到 μ 与 σ^2 的联合密度函数 $p(x | \mu, \sigma^2)$ 中的位置, 取倒伽玛分布作为 σ^2 先验分布是适当的 (见例 1.9), 取正态分布作为 μ 的先验分布也是适当的 (见例 1.5)。在考虑到 μ 与 σ^2 之间会相互影响, 故其共轭先验分布必须有乘积形式 $\pi(\mu | \sigma^2)\pi(\sigma^2)$, 其中

$$\mu | \sigma^2 \sim N(\mu_0, \sigma^2 / \kappa_0)$$

$$\sigma^2 \sim IGa(\nu_0 / 2, \nu_0 \sigma_0^2 / 2)$$

其中超参数 ν_0, μ_0, σ_0^2 在这里假设已给定, 由此容易写出 (μ, σ^2) 的联合密度函数

$$\pi(\mu, \sigma^2) \propto \sigma^{-1} (\sigma^2)^{-(\nu_0/2+1)} \exp\left\{-\frac{1}{2\sigma^2} [\nu_0 \sigma_0^2 + \kappa_0 (\mu - \mu_0)^2]\right\} \quad (1.11)$$

这种形式的分布称为正态---倒伽玛分布, 记为 $N-IGa(\nu_0, \mu_0, \sigma_0^2)$ 。

把先验密度 (1.11) 乘以正态似然, 立即可得后验密度

$$\begin{aligned}\pi(\mu, \sigma | \mathbf{x}) &\propto p(\mathbf{x} | \mu, \sigma) \pi(\mu, \sigma) \\ &\propto \sigma^{-1} (\sigma^2)^{-[(\nu_0 + n)/2 + 1]} \exp \left\{ -\frac{1}{2\sigma^2} [\nu_0 \sigma_0^2 + \kappa_0 (\mu - \mu_0)^2 + (n-1)s^2 \right. \\ &\quad \left. + n(\bar{x} - \mu)^2] \right\}\end{aligned}$$

注意到

$$\begin{aligned}&\kappa_0 (\mu - \mu_0)^2 + (\mu - \bar{x})^2 \\ &= (\kappa_0 + n) \mu^2 - 2\mu(\kappa_0 \mu_0 + n\bar{x}) + \kappa_0 \mu_0^2 + n\bar{x}^2 \\ &= (\kappa_0 + n) \left(\mu - \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + n} \right)^2 - \frac{(\kappa_0 \mu_0 + n\bar{x})^2}{\kappa_0 + n} + \kappa_0 \mu_0^2 + n\bar{x}^2 \\ &= (\kappa_0 + n) \left(\mu - \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + n} \right)^2 + \frac{n\kappa_0 (\mu_0 - \bar{x})^2}{\kappa_0 + n}\end{aligned}$$

把上式代回原式, 并记

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{x}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

(1.12)

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1) s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\mu_0 - \bar{x})^2$$

则在样本 x 给定下可得 (μ, σ^2) 的条件密度为

$$\pi(\mu, \sigma^2) \propto \sigma^{-1} (\sigma^2)^{-(\nu_n/2+1)} \exp\left\{-\frac{1}{2\sigma^2} [\nu_n \sigma_n^2 + \kappa_n (\mu - \mu_n)^2]\right\}$$

这个后验密度 (1.13) 在形式上完全与先验密度 (1.11) 相同, 只是用

ν_n, μ_n 与 $\nu_n \sigma_n^2$ 分别代替 ν_0, μ_0 与 $\nu_0 \sigma_0^2$, 故他仍是正态分布—倒

伽玛分布 $N-IGa(\nu_n, \mu_n, \sigma_n^2)$, 这说明正态—倒伽玛分布是正态均值 μ 与

正态方差 σ^2 的 (联合) 共轭先验分布。

这个后验密度 (1.13) 在形式上完全与先验密度 (1.11) 相同，只是用 ν_n, μ_n 与 $\nu_n \mu_n^2$ 分别代替 ν_0, μ_0 与 $\nu_0 \mu_0^2$ ，故他仍是正态分布—倒伽玛分布 $N-IGa(\nu_n, \mu_n, \sigma_n^2)$ ，这说明正态—倒伽玛分布是正态均值 μ 与正态方差 σ^2 的（联合）共轭先验分布。

后验分布 (1.13) 中的三个参数也可得到很好的解释，从 (1.12) 可以看出， μ_n 的先验均值 μ_0 与样本均值 \bar{x} 的加权平均，其权为 $\kappa_0 / (\kappa_0 + n)$ 和 $n / (\kappa_0 + n)$ ，这里 n 为样本容量，而 κ_0 扮演这样的角色，所提供的信息相当于“ κ_0 个样本”所提供的信息，于是 $\kappa_n = \kappa_0 + n$ 就可看作是“总样本容量”，后验自由度 ν_n 是先验自由度 ν_0 加上样本容量 n ，后验平方和 $\nu_n \mu_n^2$ 是由先验平方和 $\nu_0 \mu_0^2$ ，样本平方和 $(n-1)s^2$ 与附加的样本均值 \bar{x} 与先验均值 μ_0 之差的平方之和组成。

受联合先验分布 $\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2)$ 的启发，上述联合后验分布 $\pi(\mu, \sigma^2|x)$ 亦可分解为一个条件后验密度 $\pi(\mu|\sigma^2, x)$ 和一个边缘后验密度 $\pi(\sigma^2|x)$ 的乘积，其中

$$\begin{aligned}\mu|\sigma^2, \mathbf{x} &\sim N(\mu_n, \sigma^2/\kappa_n) \\ \sigma^2|\mathbf{x} &\sim \text{IGa}(v_n/2, v_n\sigma_n^2/2)\end{aligned}$$

把联合后验密度对 σ^2 积分可得 μ 的边缘后验密度

$$\begin{aligned}\pi(\mu|\mathbf{x}) &= \int_0^{+\infty} \pi(\mu, \sigma^2|\mathbf{x}) d\sigma^2 \\ &\propto \int_0^{+\infty} (\sigma^2)^{-(\frac{v_n+1}{2}+1)} \exp\left\{-\frac{1}{2\sigma^2}[v_n\sigma_n^2 + \kappa_n(\mu - \mu_n)^2]\right\} d\sigma^2\end{aligned}$$

利用倒伽玛密度函数的正则性，可得

$$\begin{aligned}\pi(\mu|\mathbf{x}) &\propto [v_n\sigma_n^2 + \kappa_n(\mu - \mu_n)^2]^{-\frac{v_n+1}{2}} \\ &\propto \left[1 + \frac{1}{v_n} \left(\frac{\mu - \mu_n}{\sigma_n/\sqrt{\kappa_n}}\right)^2\right]^{-\frac{v_n+1}{2}}\end{aligned}$$

这是自由度为 v_n 的 t 分布。

一般 t 分布如下密度函数

$$p(\theta|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma} \left[1 + \frac{1}{\nu} \left(\frac{\theta - \mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}, -\infty < \theta < \infty$$

其中 $\nu > 0$ 是自由度, μ 是位置参数, $\sigma > 0$ 是尺度参数, 记为 $t_\nu(\mu, \sigma^2)$,

当 $\mu = 0, \sigma = 1$ 时, $t_\nu(0, 1)$ 称为标准 t 分布, 记为 t_ν , 当 $\nu = 1$ 时, 就得哥

西分布, 一般 t 分布的期望与方差分别为

$$E(\theta) = \mu, \nu > 1$$

$$\text{Var}(\theta) = \frac{\nu}{\nu - 2} \sigma^2, \nu > 2$$

由此可见, 在本例中 μ 的边缘后验密度是 $t_{\nu_n}(\mu_n, \sigma_n^2)$, 只有在 $\nu_n > 2$ 时,

其方差存在, $\nu_n > 1$ 时, 其期望存在, 其期望为 μ_n , 方差为 $\nu_n \sigma_n^2 / (\nu_n - 2)$

例1.13 设 $x = (x_1, \dots, x_d)'$ 是 d 元随机变量，且服从 d 维正态分布， $N_d(\mu, \Sigma)$ ，这里 $\mu = (\mu_1, \dots, \mu_d)'$ 是 d 维均值向量， Σ 为其 d 阶方差-协方差矩阵，其联合密度函数为

$$p(x|\mu, \Sigma) \propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\right\}$$

若从该分布中随机抽取一样本 $y = (y_1, \dots, y_d)$ ，则此样本的联合密度函数为

$$p(y_1, \dots, y_n|\mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)' \Sigma^{-1}(y_i - \mu)\right\}$$

以下在“ Σ 是已知”的假设下来寻求均值向量 μ 的共轭先验分布，
设 μ 的共轭先验分布是 n 元正态分布，为此设

$$\mu \sim N_n(\mu_0, \Lambda_0)$$

其中均值向量 μ_0 与方差---协方差阵 Λ_0 都假设已给定，于是 μ 的后
验密度有如下形式

$$\pi(\mu|y) \propto \exp\left\{-\frac{1}{2}[(\mu - \mu_0)' \Lambda^{-1}(\mu - \mu_0) + \sum_{i=1}^n (y_i - \mu)' \Sigma^{-1}(y_i - \mu)]\right\}$$

在忽略常数因子的情况下上式可改写为

$$\begin{aligned}\pi(\mu|y) &\propto \exp\left\{-\frac{1}{2}[\mu' \Lambda_0^{-1} \mu - 2\mu' \Lambda_0^{-1} \mu_0 + n\mu' \Sigma^{-1} \mu - 2n\mu' \Sigma^{-1} \bar{y}]\right\} \\ &\propto \exp\left\{-\frac{1}{2}[\mu' (\Lambda_0^{-1} + n\Sigma^{-1}) \mu - 2\mu' (\Lambda_0^{-1} \mu_0 + n\Sigma^{-1} \bar{y})]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mu - \mu_n)' \Lambda_n^{-1} (\mu - \mu_n)\right\}\end{aligned}$$

这表明， μ 的后验密度是 n 元正态分布，其均值向量 μ_n 与方差-协方差矩阵 Λ 分别为

$$\begin{aligned}\mu_n &= (\Lambda^{-1} + n\sum^{-1})^{-1}(\Lambda^{-1}\mu_0 + n\sum^{-1}\bar{y}) \\ \Lambda_n^{-1} &= \Lambda^{-1} + n\sum^{-1}\end{aligned}$$

这个结果很类似于一元正态分布的结果，其后验均值向量是先验均值向量 μ_0 与样本均值向量 \bar{y} 的加权和，其权有先验精度矩阵 Λ_0^{-1} 与样本精度矩阵 $n\sum^{-1}$ 决定，而后验精度矩阵与样本精度矩阵之和。

1.6 充分统计量

充分统计量在简化统计问题中是非常重要的概念，也是经典统计学家和贝叶斯学者相一致的少数几个论点之一。

经典统计：设 $x = (x_1, \dots, x_n)$ 是来自分布函数 $F(x|\theta)$ 的样本， $T = T(x)$ 是统计量，假如在给定 $T(x) = t$ 的条件下， x 的条件分布与 θ 无关的话，则称该统计量为 θ 的充分统计量。

因子分解定理：一个统计量 $T(x)$ 对参数 θ 是充分的充要条件是存在一个 t 与 θ 的函数 $g(t, \theta)$ 和一个样本 x 的函数 $h(x)$ ，使得对任一样本 x 和任意 θ ，样本的联合密度 $p(x|\theta)$ 可表示为它们的乘积，即

$$p(x|\theta) = g(T(x), \theta)h(x)$$

在贝叶斯统计中，充分统计量也有一个充要条件。

定理 1.1 设 $x = (x_1, \dots, x_n)$ 是来自密度函数 $p(x|\theta)$ 的一个样本， $T = T(x)$ 是统计量，它的密度函数为 $p(t|\theta)$ ，又设 $\Pi = \{\pi(\theta)\}$ 是 θ 某个先验分布族，则 $T(x)$ 为 θ 的充分统计量的充要条件是对任一先验分布 $\pi(\theta) \in \Pi$ ，有

$$\pi(\theta|T(x)) = \pi(\theta|x)$$

即用样本分布 $p(x|\theta)$ 算得的后验分布与统计量 $T(x)$ 的分布算得的后验分布是相同的。

例1.14 设 $x = (x_1, \dots, x_n)$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 其密

度函数为

$$\begin{aligned} p(x|\mu, \sigma^2) &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} [Q + n(\bar{x} - \mu)^2]\right\} \end{aligned}$$

其中

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, Q = \sum_{i=1}^n (x_i - \bar{x})^2$$

设 $\pi(\mu, \sigma^2)$ 是任一个先验分布, 则 μ, σ^2 的后验密度为

$$\pi(\mu, \sigma^2|x) = \frac{\sigma^{-n} \pi(\mu, \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} [Q + n(\bar{x} - \mu)^2]\right\}}{\int_{-\infty}^{\infty} \int_0^{\infty} \pi \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} [Q + n(\bar{x} - \mu)^2]\right\} d\mu d\sigma^2}$$

另一方面，按经典统计，二维统计量 $T = (\bar{x}, Q)$ 恰好是 (μ, σ^2) 的充分统计量，大家知道， $\bar{x} \sim N(\mu, \sigma^2/n)$, $Q/\sigma^2 \sim \chi^2(n-1)$ ，由此可分别写出 \bar{x} 与 Q 的分布

$$p(\bar{x} | \mu, \sigma^2) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{n}{2\sigma^2} (x - \mu)^2\right\}$$

$$p(Q | \mu, \sigma^2) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)(2\sigma^2)^{\frac{n-1}{2}}} Q^{\frac{n-3}{2}} \exp\{-Q/2\sigma^2\}$$

由于 \bar{x} 与 Q 独立，所以 \bar{x} 与 Q 的联合密度容易写出

$$p(\bar{x}, Q | \mu, \sigma^2) = \frac{\sqrt{n}/\sqrt{2\pi\sigma}}{\Gamma\left(\frac{n-1}{2}\right)(2\sigma^2)^{\frac{n-1}{2}}} Q^{\frac{n-3}{2}} \exp\left\{-\frac{1}{2\sigma^2} [Q + n(\bar{x} - \mu)^2]\right\}$$

利用相同的先验分布 $\pi(\mu, \sigma^2)$ ，可得在给定 \bar{x} 与 Q 下的后验分布

$$\pi(\mu, \sigma^2 | \bar{x}, Q) = \frac{\sigma^{-n} \pi(\mu, \sigma^2) \exp\{-\frac{1}{2\sigma^2}[Q + n(\bar{x} - \mu)^2]\}}{\int_{-\infty}^{\infty} \int_0^{\infty} \pi \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}[Q + n(\bar{x} - \mu)^2]\} d\mu d\sigma^2}$$

比较这二个后验密度，可得

$$\pi(\mu, \sigma^2 | \bar{x}, Q) = \pi(\mu, \sigma^2 | x)$$

由此可见，用充分统计量 (\bar{x}, Q) 的分布算得后验分布与样本分布算得的后验分布是相同的。

二点说明,

1. 定理 1.1 给出的条件是充分必要的, 故定理 1.1 的充要条件可作为充分统计量的贝叶斯定义, 譬如在例 1.13 中把 \bar{x} 改为 x_1 , 同样可在 (x_1, Q) 给定下, 算得后验分布, 但没有上述等式, 即

$$\pi(\mu, \sigma^2 | x_1, Q) \neq \pi(\mu, \sigma^2 | x)$$

按贝叶斯定义, 统计量 (x_1, Q) 不是 (μ, σ^2) 的充分统计量。

2. 假如已知统计量 $T(x)$ 是充分的, 那么按定理 1.1, 其后验分布可用该统计量的分布算得, 由于充分统计量可简化数据, 降低维数, 故定理 1.1 亦可用来简化后验分布的计算。

例 1.15 设 $x = (x_1, \dots, x_n)$ 来自正态分布 $N = (\theta, 1)$ 的一个样本，大家知道样本均值 \bar{x} 是 θ 的充分统计量，若 θ 的先验分布取为正态分布 $N = (0, \tau^2)$ ，其中 τ^2 已知，那么 θ 的后验分布可用充分统计量 \bar{x} 的分布算得，即

$$\begin{aligned}\pi(\theta|\bar{x}) &\propto \exp\left\{-\frac{n}{2}(\bar{x} - \theta)^2 - \frac{\theta^2}{2\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}[\theta^2(n + \tau^{-2}) - 2n\theta\bar{x}]\right\} \\ &\propto \exp\left\{-\frac{n + \tau^{-2}}{2}\left(\theta - \frac{n\bar{x}}{n + \tau^{-2}}\right)^2\right\} \\ &= N\left(\frac{n\bar{x}}{n + \tau^{-2}}, \frac{1}{n + \tau^{-2}}\right)\end{aligned}$$



Homework:

PP32-33: 1, 3, 5, 6, 10, 11, 12, 15.